

기업비밀 Ⅱ급

# Multi-label Classification (NLP-side)

시니어 케어 장기기억 - 페르소나

2023. 8. 03 | 감정대화기술 PJT 윤예진



# Contents

---

A.1 Task Definition

---

A.2 Approach

---

A.3 Research Trend w/ Benchmark

---

B.1 Our Problem States

---

B.2 Experiment Plan

---

# A.1

## Task Definition

A.1	Task Definition
-----	-----------------

A.2	Approach
-----	----------

A.3	Research Trend w/ Benchmark
-----	-----------------------------

B.1	Our Problem States
-----	--------------------

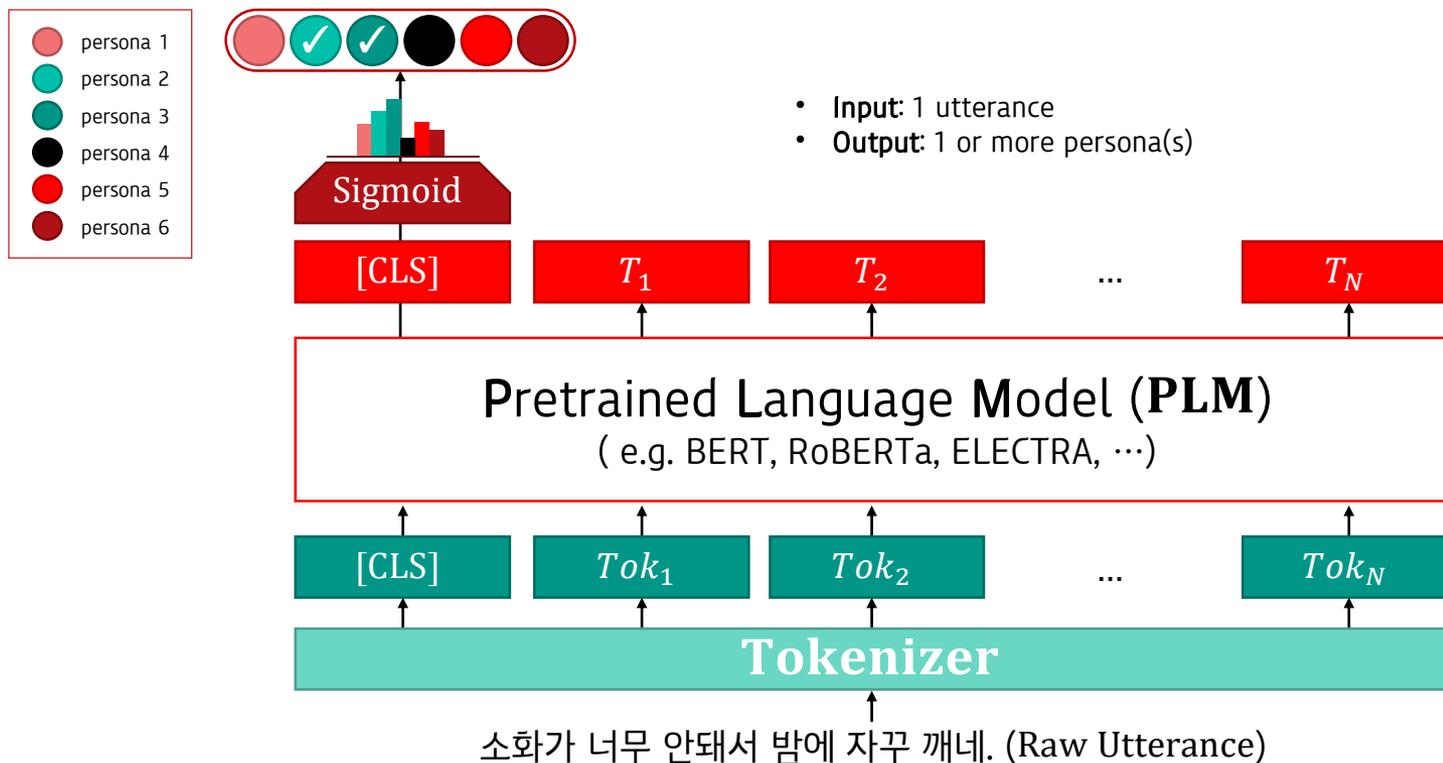
B.2	Experiment Plan
-----	-----------------

# 00 Multi-label Classification :: NLP-side

## A.1 Task Definition

### • Multi-label Classification w/ Deep Learning

- 전통적인 machine learning 문제에서 정의하는 classification의 한 종류
- 각 instance에 label이 여러 개(multiple label) 할당될 수 있는 classification task



## A.2

# Approach

- \* OvR: One-versus-Rest
- \* MLC with threshold
- \* MLC without threshold
  - Top-K: Select top-k label
- \* Generative Classification

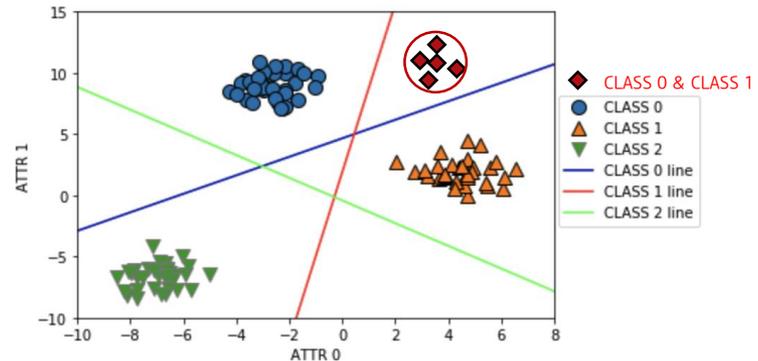
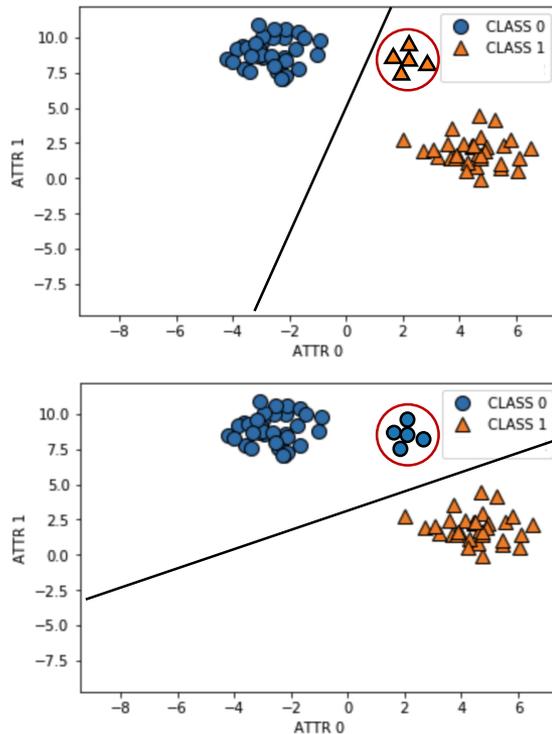
A.1	Task Definition
A.2	Approach
A.3	Research Trend w/ Benchmark
B.1	Our Problem States
B.2	Experiment Plan

# 01 OvR

## A.2 Approach

- **One-vs-Rest Method**

- A heuristic method for using binary classification for multi-class classification
- Label 수 만큼의 이진 분류 모델을 학습, 1회의 추론에서 모든 이진 분류 모델을 사용



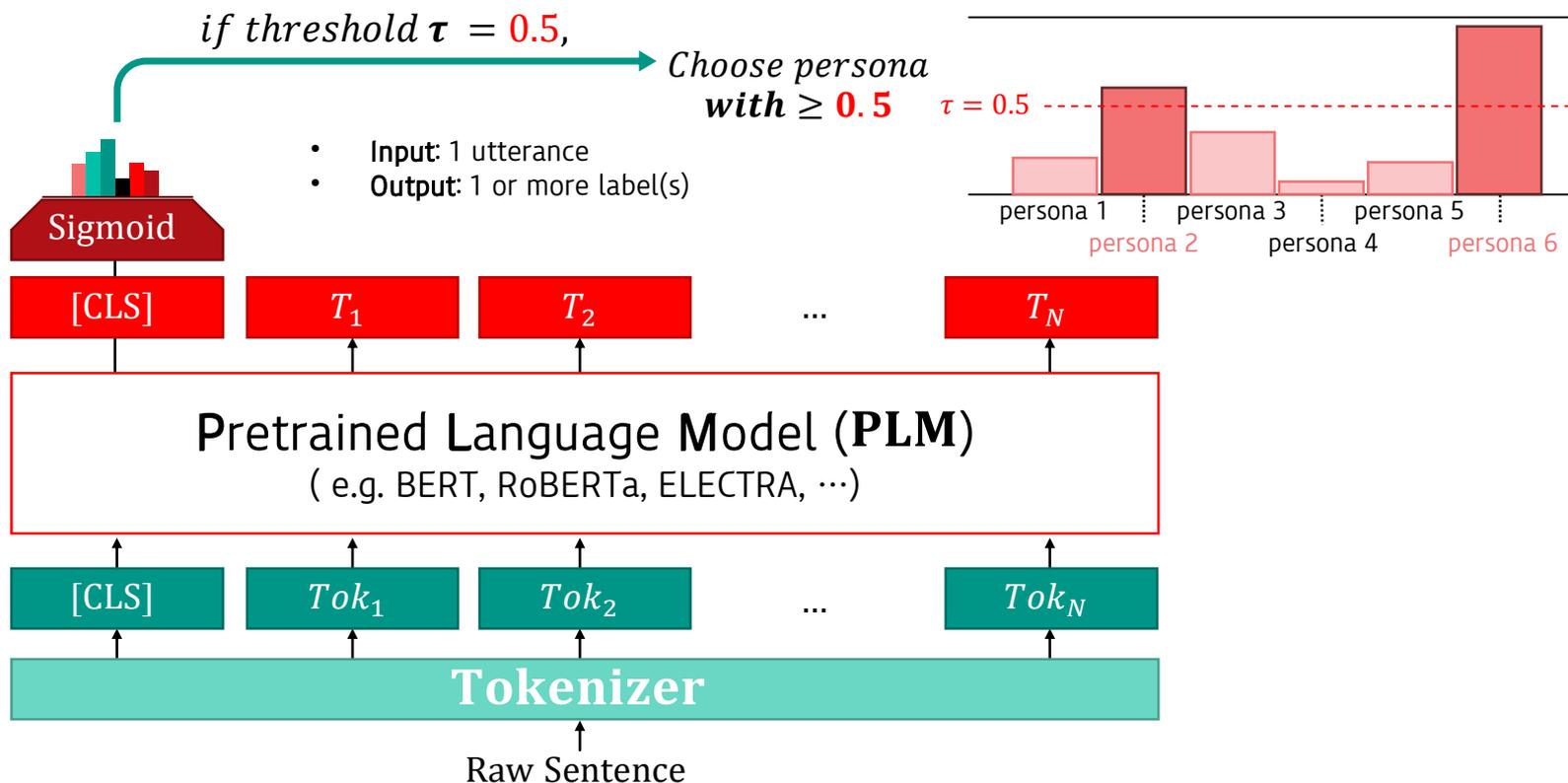
Multi-label classification using OvR

## 02 MLC w/Threshold

### A.2 Approach

- Apply Sigmoid for multi-label classification problem

- classification threshold은 default로 0.5로 선정 ( $\tau$ : threshold, hyperparameter)

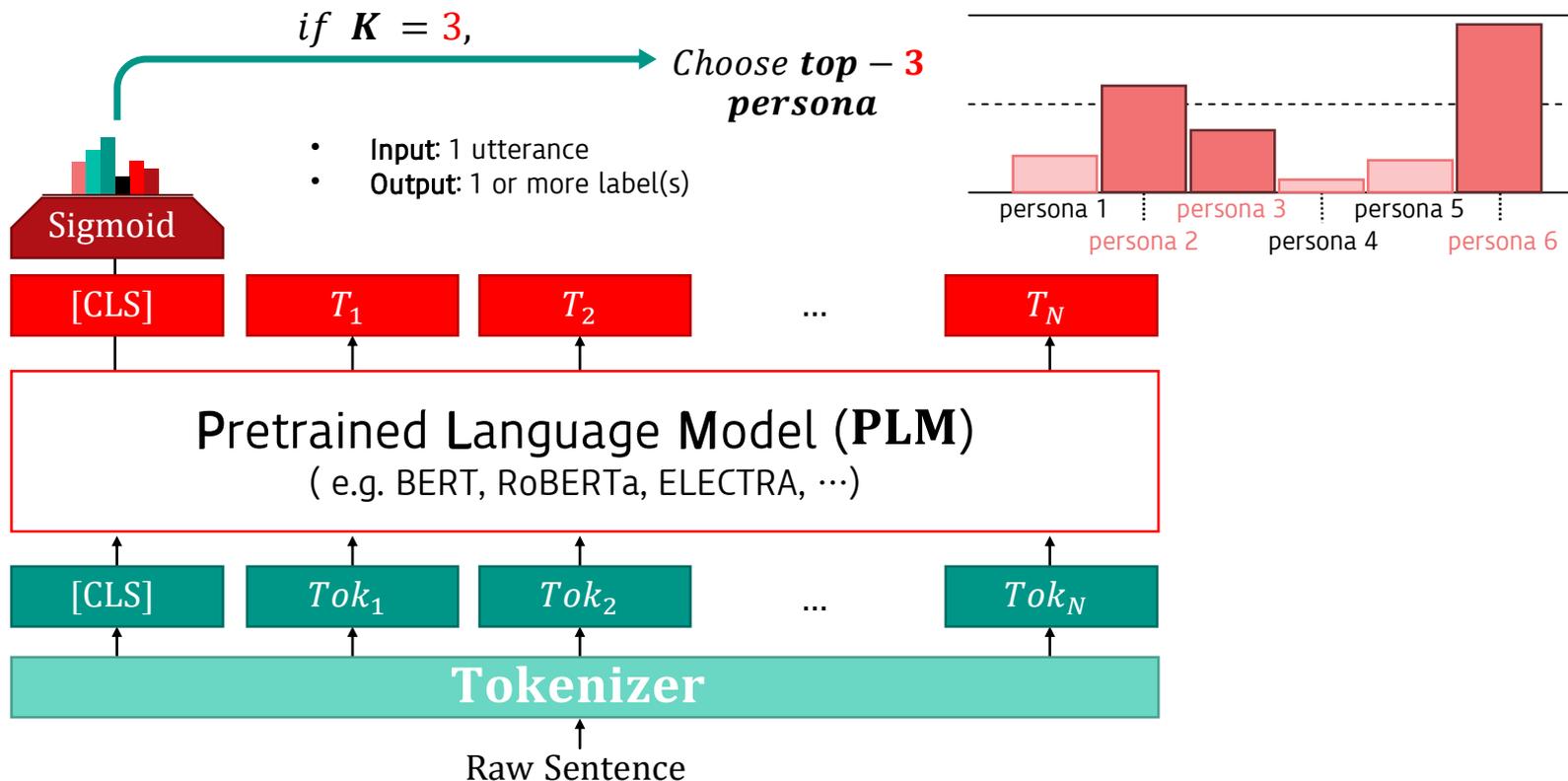


# 03 MLC w/o Threshold :: top-K

## A.2 Approach

- Apply Sigmoid for multi-label classification problem

- top-K개의 intent 선택 ( $K$ : hyperparameter)

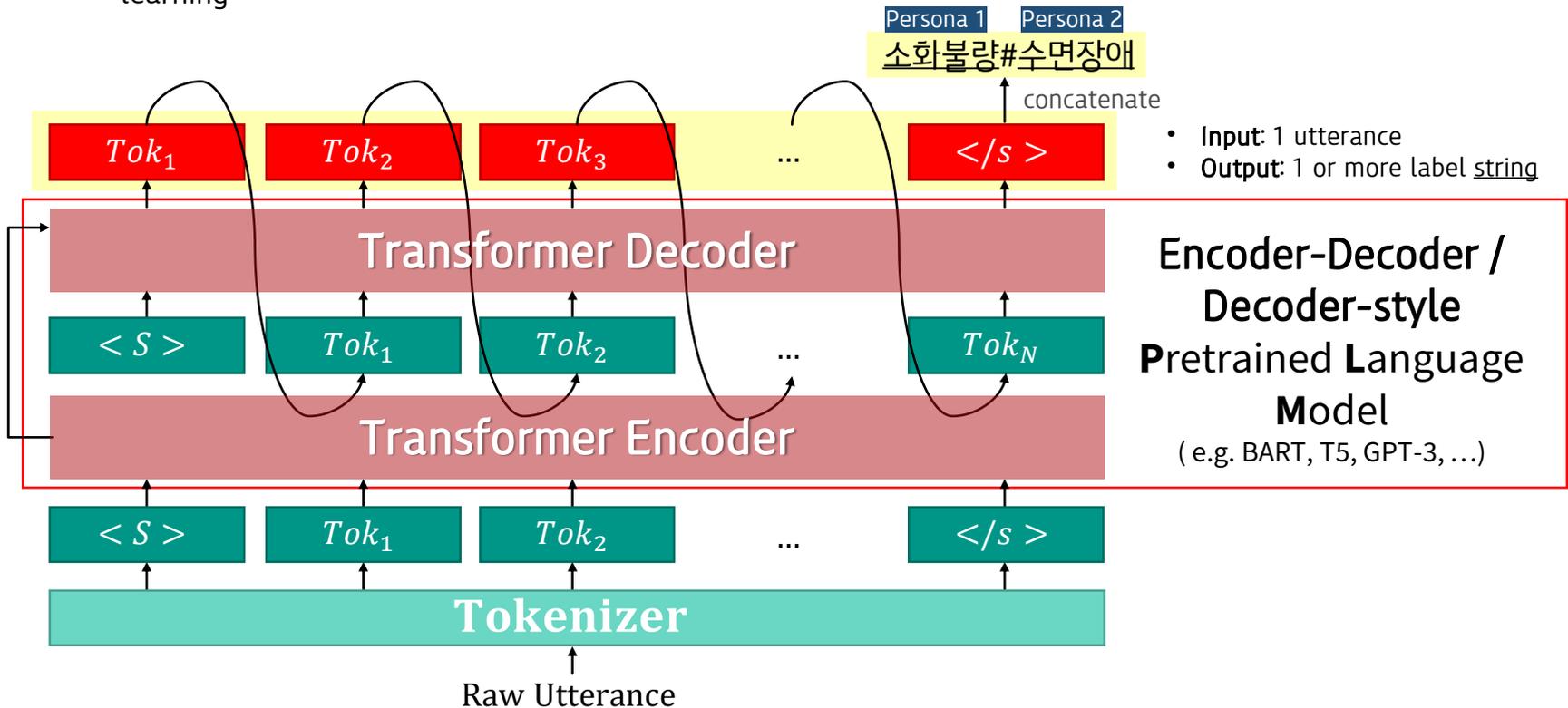


# 04 Generative Classification

## A.2 Approach

- Apply text-to-text framework

- Decoder-style 모델이 input으로 받은 utterance에 대해 output으로 분류할 label 조합을 출력하도록 transfer learning



## A.3

# Research Trend (NLP)

\* MLC : **M**ulti-**L**abel **C**lassification

- **Hierarchical Label**
  - WoS
  - 20 Newsgroups
- **Non-hierarchical Label**
  - Reuter-21578

\* XMC : **E**xtr<sup>e</sup>m<sup>e</sup> **M**ulti-**L**abel **C**lassification

- Amazon-670K

A.1	Task Definition
A.2	Approach
A.3	Research Trend w/ Benchmark
B.1	Our Problem States
B.2	Experiment Plan

# 01 MLC - Hierarchical :: WoS

## A.3 Research Trends w/ Benchmarks

- **WoS (Introduced by Kowsari et al. 2017)**

- Web of Science, Multi- & hierarchical- label document classification dataset
- Size: Total 46,985 documents
- Label: 134 Area (up to 2-label)

label ▶

Domain	Number of Document	Number of Area
Biochemistry	5,687	9
Civil Engineering	4,237	11
Computer Science	6,514	17
Electrical Engineering	5,483	16
Medical Sciences	14,625	53
Mechanical Engineering	3,297	9
Psychology	7,142	19
<b>Total</b>	<b>46,985</b>	<b>134</b>

Data Set	Training	Testing	Level 1	Level 2
WOS-11967	8018	3949	7	35
WOS-46985	31479	15506	7	134
WOS-5736	4588	1148	3	11

- Repository: [data.mendeley.com/datasets/9rw3vkcfy4/6](https://data.mendeley.com/datasets/9rw3vkcfy4/6)
- Leaderboard: [paperswithcode.com/dataset/web-of-science-dataset/](https://paperswithcode.com/dataset/web-of-science-dataset/)
  - WOS - 5736 : 91.28 (Accuracy)
  - WOS-11967 : 86.07 (Accuracy)
  - WOS-46985 : 76.58 (Accuracy)

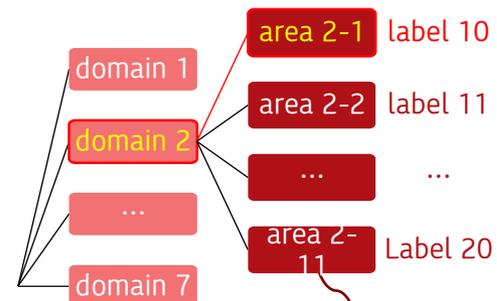
# 01 MLC - Hierarchical :: WoS

## A.3 Research Trends w/ Benchmarks

- **WoS** (Introduced by Kowsari et al. 2017)

- Web of Science, Multi- & hierarchical- label document classification dataset
- Size: Total 46,985 documents
- Label: 134 Area (up to 2-label)

label ▶	Domain	Number of Document	Number of Area
	Biochemistry	5,687	9
	Civil Engineering	4,237	11
	Computer Science	6,514	17
	Electrical Engineering	5,483	16
	Medical Sciences	14,625	53
	Mechanical Engineering	3,297	9
	Psychology	7,142	19
	<b>Total</b>	<b>46,985</b>	<b>134</b>



Data Set	Training	Testing	Level 1	Level 2
WOS-11967	8018	3949	7	35
WOS-46985	31479	15506	7	134
WOS-5736	4588	1148	3	11

sample ▼

Document (input)	label (int32)	label_level_1 (int32)	label_level_2 (int32)
input_data (string) "The aim of this study was to investigate (a) the behavioral cues that are displayed by, and trait judgments..."	10	2	2
"The detection of negative emotions through daily activities such as writing and drawing is useful for..."	1	0	1
"DNA/RNA sequencing has recently become a primary way researchers generate biological data for further analysis..."	29	6	1

## 02 MLC - Hierarchical :: 20 Newsgroups

### A.3 Research Trends w/ Benchmarks

- **20 Newsgroups** (Introduced by Lang et al. 1995)

- News group documents, partitioned evenly across 20 different newsgroups
- Size: 18,846 documents
- Label: 20 topics (up to 5-label)

label ►

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast
sci.crypt sci.electronics sci.med sci.space	talk.religion.misc alt.atheism soc.religion.christian

- Repository: [/~jason/20Newsgroups/](#)
- Leaderboard: [paperswithcode.com/dataset/20-newsgroups](#)
  - Text classification : 89.5 (Accuracy)
  - Text clustering : 41.25 (Accuracy)

#### sample ▼

```
Newsgroup: sci.med → target label
document_id: 57110
From: bed@intacc.uucp (Deb Waddington) → meta info.
Subject: INFO NEEDED: Gaucher's Disease

...
Gaucher's disease symptoms include: brittle bones (he lost 9
inches off his hieght); enlarged liver and spleen; internal
bleeding; and fatigue (all the time). The problem (in Type 1) is
attributed to a genetic mutation where there is a lack of the
enzyme glucocerebroside in macrophages so the cells swell up.
This will eventually cause death.
...
NEED INFO:
I have researched Gaucher's disease at the library but am relying
on netlanders to provide me with any additional information:
...
**Basically ANY HELP YOU CAN OFFER

Thanks so very much!

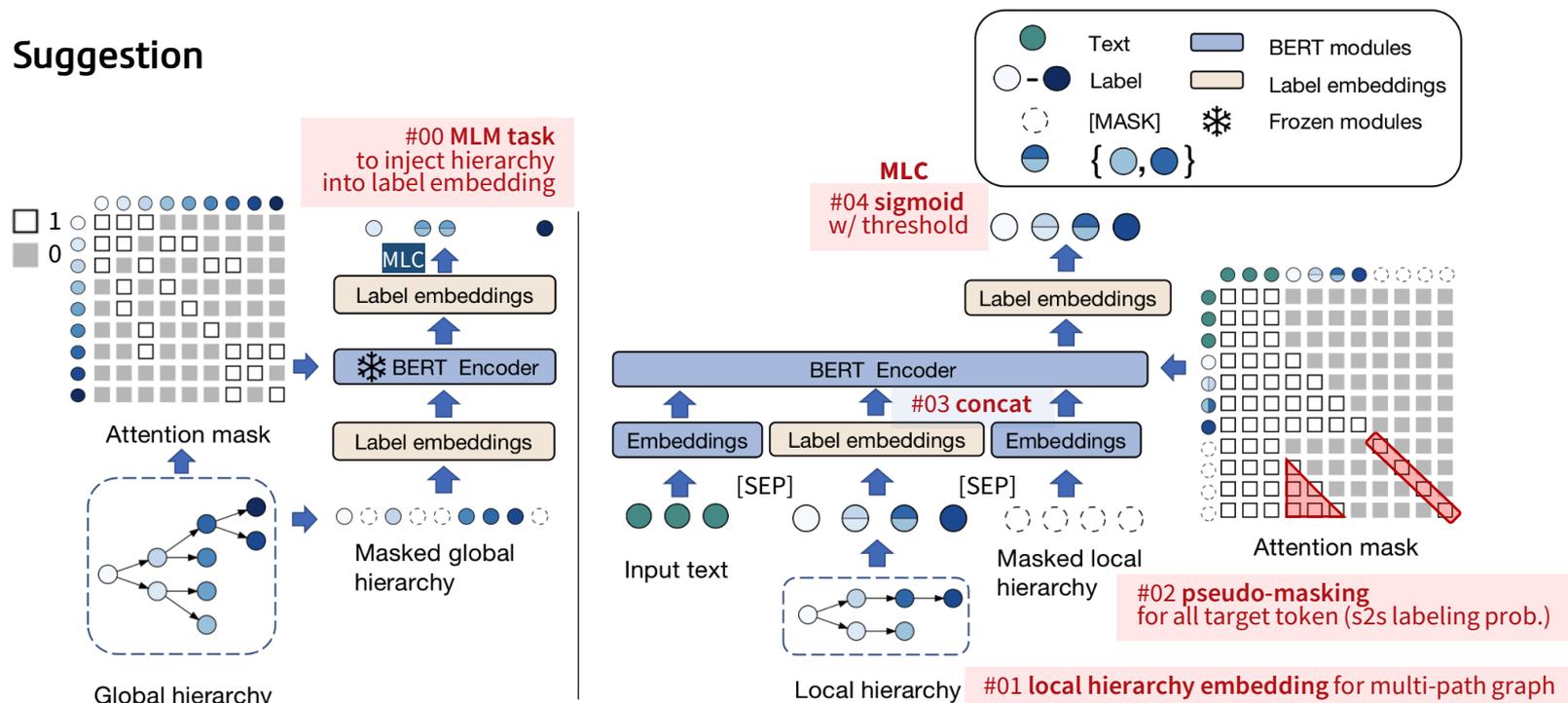
Deborah
```

# 03 MLC - Hierarchical :: Related Work

## A.3 Research Trends w/ Benchmarks

■ HBGL) Jiang et al. Exploiting Global and Local Hierarchies for Hierarchical Text Classification, EMNLP, 2022

### • Suggestion



- Hierarchy-guided BERT w/ Global & Local hierarchies\* s2s-ft: seq2seq finetuning
- Model avoids the intentional fusion of semantic and hierarchical modules by directly modeling semantic and hierarchical info. from BERT
- w/o graph encoder which is static and models the same graph redundantly

# 04 MLC - Non-hierarchical :: Reuters-21578

## A.3 Research Trends w/ Benchmarks

- **Reuters-21578** (Introduced by Lewis et al. 1987)

- A collection of news article dataset
- Size: Total 21,578 documents
- Label: 135 topics (up to 16-label)

imbalanced multi-label ▼

#Topics	# train	# test	# other	Total	#Topics	# train	# test	# other	Total
0	1,828	280	8,103	10,211	8	4	2	-	6
1	6,552	2,581	361	9,494	9	4	2	-	6
2	890	309	135	1,334	10	3	1	-	4
3	191	64	55	310	11	-	1	1	2
4	62	32	10	104	12	1	1	-	2
5	39	14	8	61	13	-	-	-	-
6	21	6	3	30	14	-	2	-	2
7	7	4	-	11	15	-	-	-	-
					16	1	-	-	1

sample ▼

Multi-label

Topic	Text
'grain', 'wheat', 'corn', 'barley', 'oat', 'sorghum'	The U.S. Agriculture Department reported the farmer-owned reserve national five-day average price through February 25 as follows ...
'veg-oil', 'linseed', 'lin-oil', 'soy-oil', 'sun-oil', 'soybean', 'oilseed', ...	Argentine grain board figures show crop registrations of grains, oilseeds and their products to February 11, in thousands of tonnes, ...

Input text

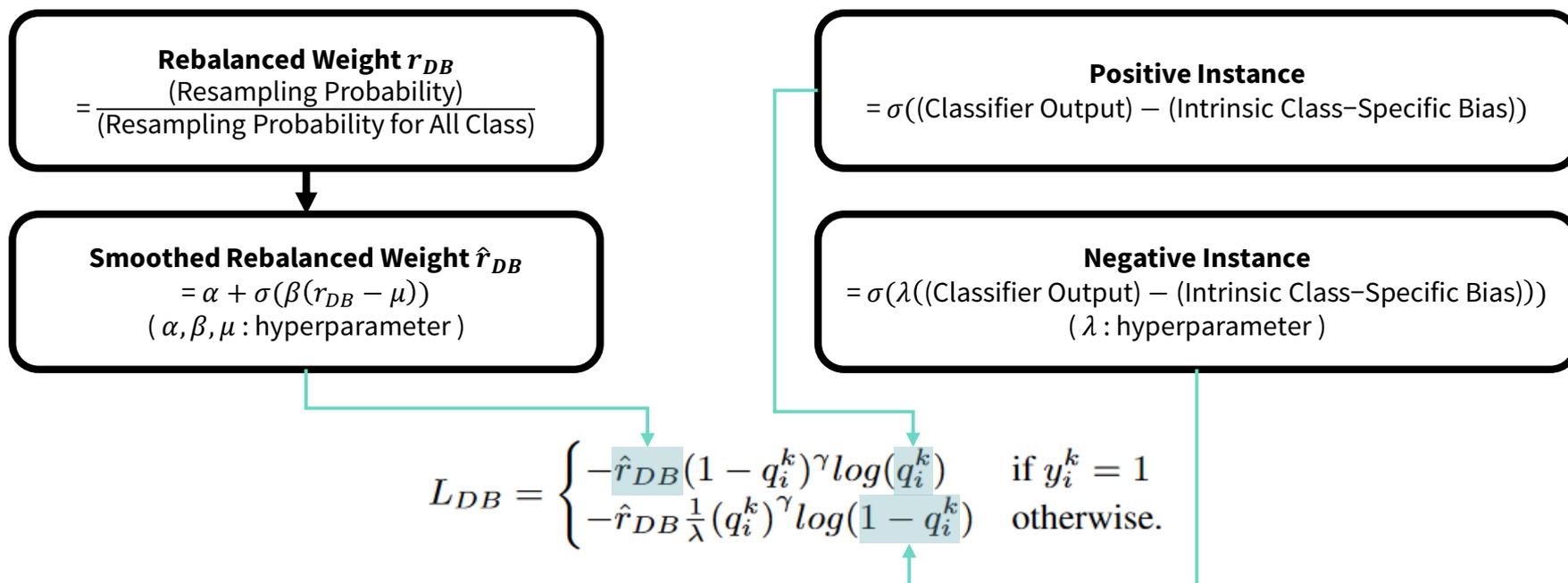
- Repository: [/testcollections/reuters21578/](#)
- Leaderboard: [paperswithcode.com/dataset/reuters-21578/](#)
  - Document Classification : 97.44 (Accuracy)
  - Multi-label Text Classification : 90.74 (Accuracy)

## 05 MLC - Non-hierarchical :: Related Work

### A.3 Research Trends w/ Benchmarks

■ BalancedLossNLP) Huang et al. Balancing Methods for Multi-label Text Classification with Long-Tailed Class Distribution, EMNLP, 2021

- **Suggestion**



- Apply distribution-balanced loss function to train, to resolve the long-tailed class imbalance for multi-label classification

# 06 XLC :: Amazon-670K

## A.3 Research Trends w/ Benchmarks

- Amazon-670K

- Task definition: find most relevant labels from an extreme large label set for an input text instance

label ▼			Total # of label		Average # of label for each instance	
Dataset	$N_{train}$	$N_{test}$	$L$	$\bar{L}$	$\bar{n}$	$d_{tfidf}$
Eurlex-4K	15,449	3,865	3,956	5.30	20.79	186,104
Wiki10-31K	14,146	6,616	30,938	18.64	8.52	101,938
AmazonCat-13K	1,186,239	306,782	13,330	5.04	448.57	203,882
Wiki-500K	1,779,881	769,421	501,070	4.75	16.86	2,381,304
Amazon-670K	490,449	153,025	670,091	5.45	3.99	135,909
Amazon-3M	1,717,899	742,507	2,812,281	36.04	22.02	337,067

e.g. Amazon-670K

sample ▼

```
{
  "uid": "0816614369",
  "title": "Kant's Critical Philosophy: ...",
  "content": "Text: English, French ...",
  "target_ind": [27024, 27040, 27079, ...],
  "target_rel": [1.0, 1.0, 1.0, 1.0, 1.0, ...]
}
```

Document (input)

```
Id: 192731
ASIN: 0816614369
title: Kant's Critical Philosophy: ...
group: Book
salesrank: 290384
similar: 5 0231068131 0816643415 0942299515 1890951242 0816616752
categories: 2
|Books[283155]|Subjects[1000]|Nonfiction[53]
|Philosophy[11019]|Epistemology[11035]
|Books[283155]|Subjects[1000]|Nonfiction[53]
|Philosophy[11019]|History, 17th & 18th Century[11051]
reviews: total: 2 downloaded: 2 avg rating: 5
2000-8-10 cutomer: A2UUWETDYA2EKV rating: 5 votes: 15 helpful: 14
2001-2-15 cutomer: A1R6E5S8WLUXEX rating: 5 votes: 13 helpful: 9
```

Target multi-label

## A.3.1

# Research Trend (Vision)

\* PLC : Partial Label Classification

- Observed vs. Unobserved Annotation

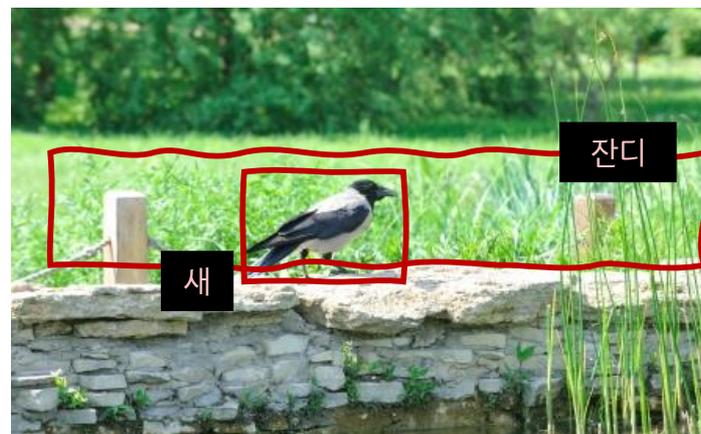
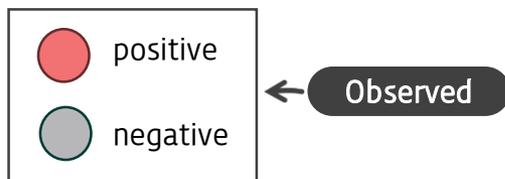
A.1	Task Definition
A.2	Approach
A.3	Research Trend w/ Benchmark
B.1	Our Problem States
B.2	Experiment Plan

# 07 Partial Label Classification

## A.3 Research Trends w/ Benchmarks

- **Partial Label Classification**

- Observed vs. Unobserved Annotation



Ground Truth



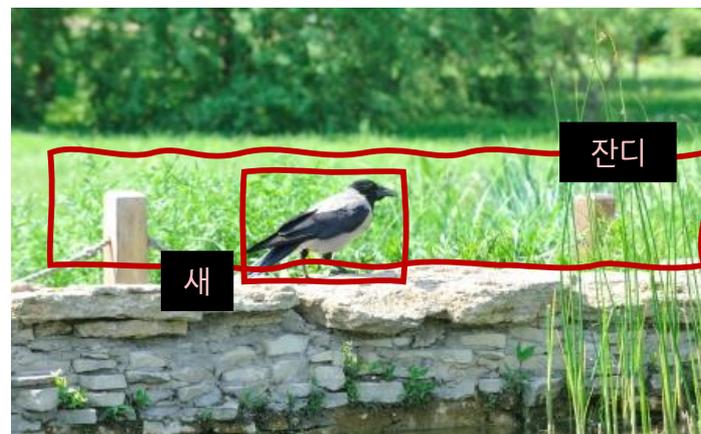
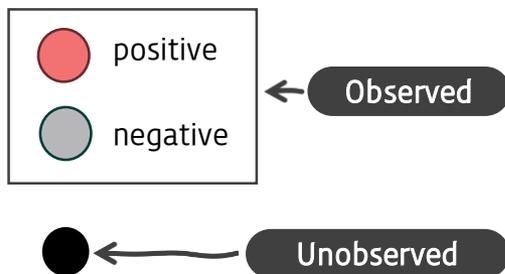
- 새
- 잔디
- 하늘
- 악어
- 호랑이
- 장미

# 07 Partial Label Classification

## A.3 Research Trends w/ Benchmarks

- Partial Label Classification

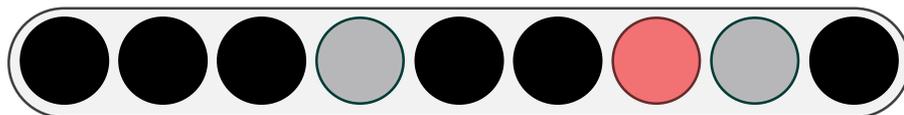
- Observed vs. Unobserved Annotation



Ground Truth



Observed Label



- 새
- 잔디
- 하늘
- 악어
- 호랑이
- 장미

# 08 Partial Label Classification :: Related Work

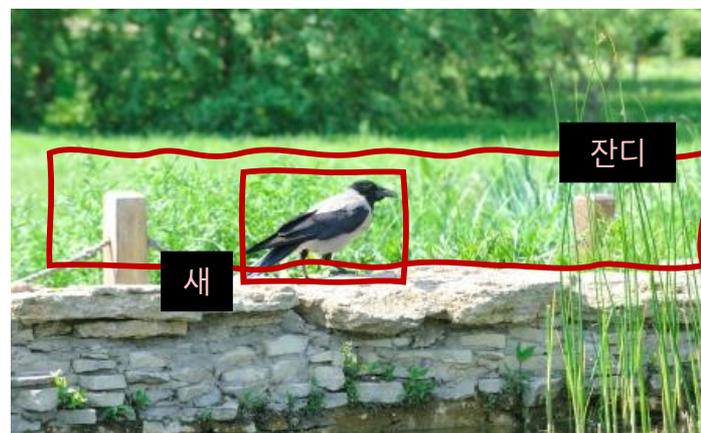
## A.3 Research Trends w/ Benchmarks

■ Zhang, Xin, et al. An Effective Approach for Multi-label Classification with Missing Labels, IEEE, 2022

### • Suggestion

DIFFERENT MISSING-LABEL SETTINGS. ✓, ×, ∅ REPRESENT THAT CURRENT INSTANCE BELONGS TO THIS CLASS, DOES NOT BELONG TO THIS CLASS, AND LACKS RELATED LABEL, RESPECTIVELY.

Settings	Class 1	Class 2	Class 3	Class 4	Class 5
FOL	✓	×	✓	×	✓
POL	✓	∅	∅	×	✓
PPL	✓	∅	✓	∅	∅
SPL	✓	∅	∅	∅	∅



class	새	잔디	하늘	악어	호랑이	장미	사람
POL	●	●	○	●	●	○	○
PPL	●	●		●	●		

← “Missing Label”

- POL: partially observed labels
  - 일부 레이블이 관측되지 않았는데, 적어도 관측된 레이블은 확실히 +인지 -인지 알고 있음
- PPL: partially observed positive labels
  - 일부 레이블이 관측되지 않았는데, 관측된 레이블이 확실히 +인 것만 알고 있음

# 08 Partial Label Classification :: Related Work

## A.3 Research Trends w/ Benchmarks

■ Zhang, Xin, et al. An Effective Approach for Multi-label Classification with Missing Labels, IEEE, 2022

- **Suggestion**

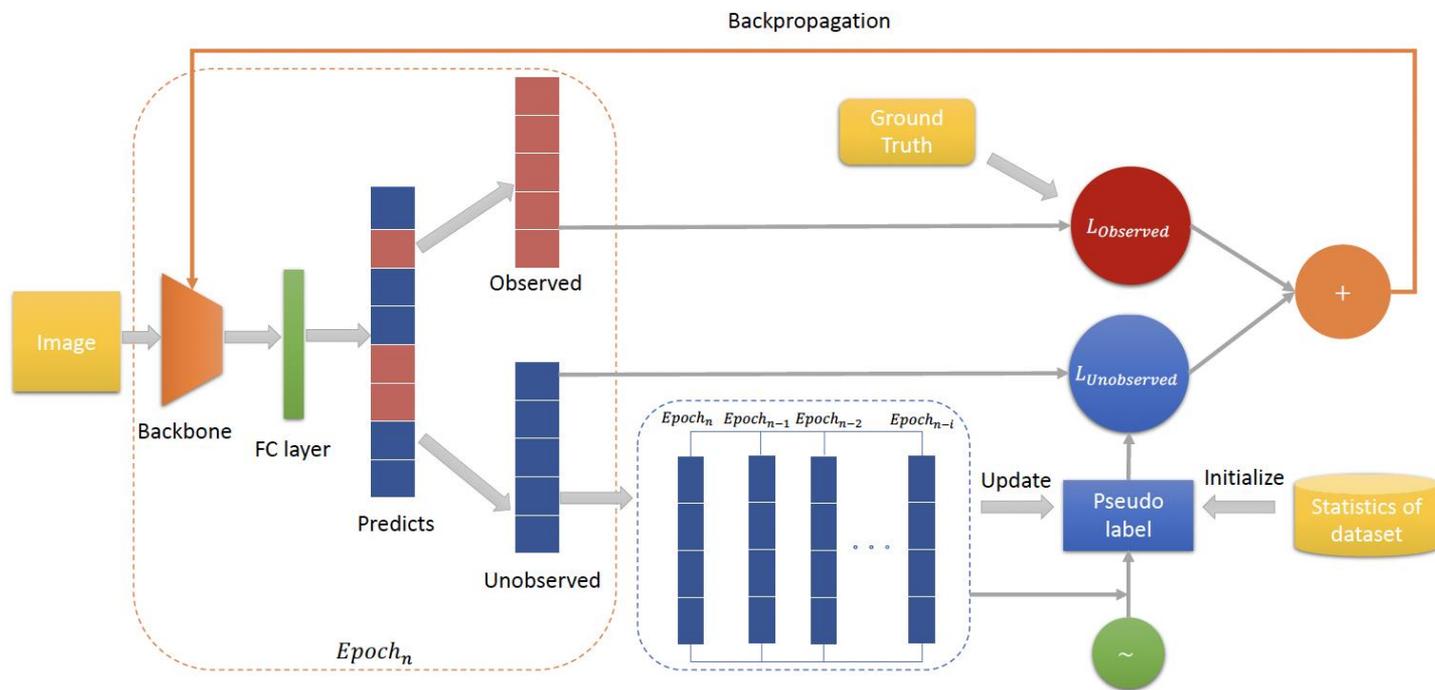


Fig. 1. Pipeline of our method. We feed an image to the classifier network and obtain the predictions. Then different strategies are utilized to deal with observed and unobserved labels. For the unobserved part, we introduce pseudo labels by leveraging the prior knowledge of the dataset and design a novel loss function  $\mathcal{L}_{Unobserved}$ . For the observed part, we treat it as a full-label classification problem. The total loss is defined by adding the losses of these two parts with designed weights. The network used in our method is the most common classification network, which means that there is no additional complexity in the network structure.

## 08 Partial Label Classification :: Related Work

### A.3 Research Trends w/ Benchmarks

■ Zhang, Xin, et al. An Effective Approach for Multi-label Classification with Missing Labels, IEEE, 2022

#### • Suggestion

- POL: partially observed labels
  - 일부 레이블이 관측되지 않았는데, 적어도 관측된 레이블은 확실히 +인지 -인지 알고 있음

class	새	잔디	하늘	악어	호랑이	장미	사람
POL	●	●	○	●	●	○	○

Unobserved label

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{Unobserved}}}_{\downarrow} + \mathcal{L}_{\text{Observed}}$$

: Same as BCE Loss

1) initialization: pseudo-label

$$\tilde{y}_{ij} = \begin{cases} \min\left(\frac{\max(\frac{P_p}{m}, 1) - P_j}{T_{u_i}}, 1\right) & \max(\frac{P_p}{m}, 1) > P_i \\ 0, & \textit{otherwise} \end{cases}$$

2) update pseudo-label (during training)

# B.1

## Our Problem States

A.1	Task Definition
A.2	Approach
A.3	Research Trend w/ Benchmark
<b>B.1</b>	<b>Our Problem States</b>
B.2	Experiment Plan

# 09 Problem States

## B.1 Our Problem States

### • 시니어 케어 장기기억 - 페르소나

SID	UID	Speaker	Utterance	persona/utt.	persona/sess.
1	1	시스템	\$(SeniorName)님. 지난밤 잠은 꼭 주무셨어요? 컨디션 좋은 하루를 시작하셨길 바랍니다.	[분류없음]	수면장애
1	2	사용자	요즘 계속 잠을 못 자서 컨디션이 꺾이야.	수면장애+	
1	3	시스템	수면을 취하셔야 피로가 풀리실 텐데 걱정이네요. 혹시 잠을 못 주무시는 특별한 이유가 있으신가요?	수면장애+	
1	4	사용자	코막힘 때문에 잘 때 숨쉬기가 불편해서 그래.	수면장애+감기 의심	
1	5	시스템	코막힘 증상이 있으면 주무실 때 신경이 많이 쓰일 것 같아요. 감기 때문에 코막힘 증상이 생기셨나요?	감기 의심	
1	6	사용자	감기가 아니라 비염이 심해진 것 같아.	[분류없음]	
1	7	시스템	\$(SeniorName)님께서 비염이 있으시군요. 비염 치료를 위해 병원에는 방문해 보셨어요?	[분류없음]	
1	8	사용자	밤마다 너무 괴로워서 오늘 가보려고 해.	[분류없음]	
1	9	시스템	좋은 생각이에요. 치료를 받으시면 밤에 한결 편안하게 주무실 수 있을 거예요.	[분류없음]	
1	10	사용자	스트레스로 과식까지 하게 돼서 꼭 낮고 싶어.	[분류없음]	
1	11	시스템	\$(SeniorName)님께서 그동안 스트레스를 많이 받고 계셨나 보네요. 비염 증상이 하루빨리 나아지시기를 바랍니다.	[분류없음]	

#### – Problem States: Multi-persona Classification for utterance

- Class-imbalanced: NA vs. others
  - 질병: 19+1 (NA) & state (+/-/의심)
  - 일상: 17+1 (NA) & state (+/-)

Train	Test	Utterance	Session
Utterance		85.19	21.36
Session		7.56	33.33

1. Session단위의 prediction을 missing label setting으로 간주하고 utterance 단위의 prediction으로 확장
2. Generative Classification 성능 개선
3. Encoder-based Classification 적용

## B.2

# Experiment Plan

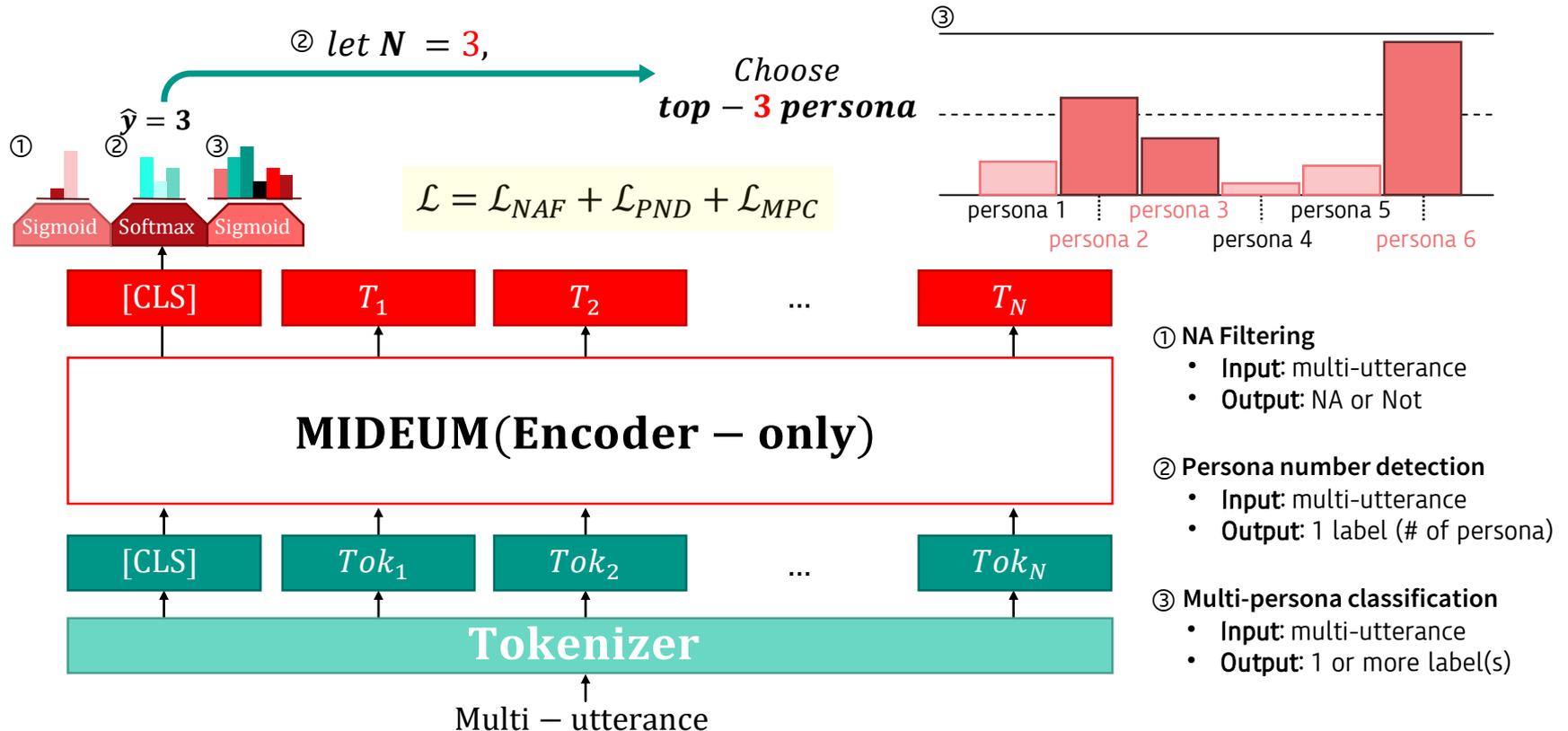
A.1	Task Definition
A.2	Approach
A.3	Research Trend w/ Benchmark
B.1	Our Problem States
<b>B.2</b>	<b>Experiment Plan</b>

# 10 MLC w/joint-learning

## B.2 Experiment Plan

- **Multi-label Classification w/ joint-loss**

- Backbone: KT MIDEUM (kt-ulm-base) encoder-only





# 11 MLC w/joint-learning

## B.2 Experiment Plan

- **Future Works**

- Generative method로 확장
  - 필요에 따라 Decoder를 포함한 backbone모델로 변경
- Embedding 설계
  - 화자 정보
  - 세션 정보
  - ...

**kt**